Auf das Design kommt es an: experimentelle Befunde zu komplexen Settings in Faktoriellen Surveys

Auspurg, Katrin; Hinz, Thomas; Liebig, Stefan; Sauer, Carsten

Veröffentlichungsversion / Published Version
Zeitschriftenartikel / journal article

Auf das Design kommt es an
Experimentelle Befunde zu komplexen Settings in Faktoriellen Surveys

Katrin Auspurg, Thomas Hinz, Stefan Liebig und Carsten Sauer

1 Faktorielle Surveys: Prinzip und Forschungsstand

In faktoriellen Surveys werden den Befragten hypothetische Objekt- oder Situationsbeschreibungen (Vignetten) zur Bewertung vorgelegt, in denen Merkmale (Dimensionen) experimentell in ihren Ausprägungen (Levels) variiert werden (allgemeine Einführungen bieten Rossi/Anderson 1982; Jasso 2006; Beck/Opp 2001). Abbildung 1 zeigt eine Vignette, wie sie zur Bestimmung von Determinanten der Einkommensgerechtigkeit, einem klassischen Anwendungsgebiet von Vignetten, eingesetzt werden kann. Ein fiktiver Einkommensbezieher wird anhand der fünf Merkmale Geschlecht, Alter, Ausbildungsabschluss, Beruf und Bruttoeinkommen beschrieben. Aufgabe der Befragten ist es, auf Basis dieser Angaben die Gerechtigkeit des Einkommens zu beurteilen.

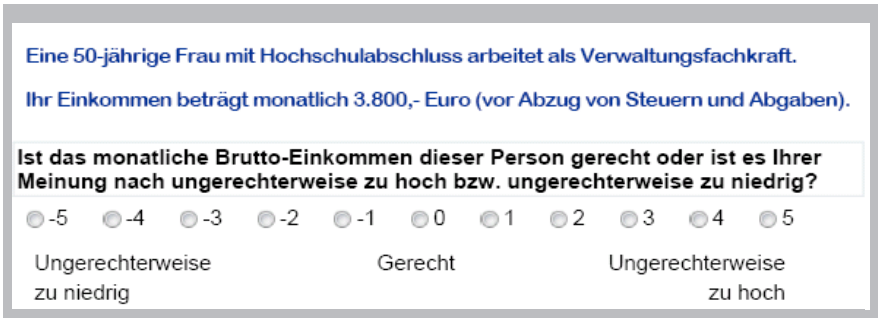

Abbildung 1: Beispielvignette

Indem die Ausprägungen dieser fünf Merkmale über die Vignetten hinweg unabhängig voneinander variieren, lässt sich ihre jeweilige Bedeutung für die Gerechtigkeitsurteile bestimmen. Ist etwa das Geschlecht der beschriebenen Person für den Befragten urteilsrelevant? Aufgrund der experimentellen Variation unterscheiden sich die beschriebenen weiblichen und männlichen Einkommensbezieher insgesamt nicht in ihren Eigenschaften. Dies bietet einen wesentlichen Vorteil gegenüber „realen“ Arbeitsmarktdaten, in denen das Geschlecht stark mit den Berufssparten, dem Ausmaß der Berufserfahrung und weiteren einkommensrelevanten Merkmalen korreliert und daher weibliche und männliche Einkommensbezieher nur sehr bedingt miteinander vergleichbar sind. Werden Männern höhere Einkommen als Frauen zugestanden, ist mit nicht-experimentellen Daten nicht verlässlich feststellbar, inwieweit es sich um eine Einkommensdiskriminierung von Frauen handelt oder aber die Lohndifferenzen nicht kontrollierten Ausstattungs- und Leistungsmerkmalen geschuldet sind (Lorenz 1993).

Neben dieser gegenüber „herkömmlichen“ Umfragedaten weitaus besseren Möglichkeit, kausale Zusammenhänge aufzuklären, gilt als ein weiterer Vorteil, dass die Befragten mit mehreren Informationen simultan konfrontiert werden. Dies erzeugt gegenüber einfachen Itemabfragen eine bessere Korrespondenz zu realen Entscheidungs- und Bewertungssituationen, in denen in der Regel auch mehrere Merkmale zusammentreten. Überdies zwingt dies die Befragten dazu, die einzelnen Dimensionen gegeneinander abzuwägen. Es kann nicht alles gleichermaßen als sehr wichtig erachtet werden, da die Merkmale in ein *gemeinsames* Urteil zu integrieren sind, womit besserer Aufschluss über die (relative) Bedeutung der einzelnen Merkmale gewonnen wird. Schließlich lassen sich im Gegensatz zur klassischen Laborforschung auch Stichproben der Allgemeinbevölkerung ohne großen Aufwand realisieren.

Inhaltlich konzentrieren sich die Anwendungen faktorieller Surveys auf die Einstellungs- und Normenmessung (z.B. Bestimmung gerechter Einkommenshöhen und Strafmaße), es finden sich aber ebenfalls Abfragen von Entscheidungen und Handlungsintentionen (z.B. Umzugs- und Kaufentscheidungen) oder Messungen des Status von Personen und Haushalten (für einen Überblick: Auspurg et al. 2009a; Wallander 2009). Die Methode zählt mittlerweile zum Repertoire vieler soziologischer Teildisziplinen. Die breite Einsetzbarkeit zeigt sich auch an den über 100 Veröffentlichungen, die ein aktueller Review-Artikel seit den ersten Anleitungen von Rossi im Jahr 1982 bis zum Jahr 2006 für die englischsprachigen *core* oder *priority* Zeitschriften der Soziologie ausweist (Wallander 2009).

Im Kontrast zu dieser häufigen Anwendung steht die geringe Erforschung des Verfahrens (für Ausnahmen zur Stichprobenziehung von Vignetten Steiner/Atzmüller 2006; Dülmer 2007). Dabei werden von Beginn an diverse methodische Probleme diskutiert. Spekuliert wird etwa darüber, ob und ab welcher Anzahl an Dimensionen eine Überforderung der Befragten eintritt und wie sich diese äußert. Ebenso ist unklar, inwieweit das Standardvorgehen, den einzelnen Befragten gleich mehrere Vignetten zur Beurteilung vorzulegen, zu Ermüdungseffekten und damit inkonsistenten Antworten sowie ungewünschten Ausstrahlungseffekten der anfänglichen Urteile auf die späteren führt. Schließlich werden invalide Urteile auch deshalb befürchtet, weil einzelne Merkmale aufgrund einer auffälligeren Operationalisierung ein stärkeres Einflussgewicht erhalten könnten (Wason et al. 2002; Wittink et al. 1990). Allgemein ist unklar, wie die Datenqualität faktorieller Surveys einzuschätzen und zu verbessern ist.

Das in vorliegendem Beitrag vorgestellte Forschungsprojekt zielt genau in diese Lücke in der Methodenforschung. Es handelt sich um das von der Deutschen Forschungsgemeinschaft (DFG) geförderte Projekt „Der faktorielle Survey als Instrument zur Einstellungsmessung in Umfragen“. Zentrale Forschungsfragen und Erhebungstechniken werden im Folgenden dargestellt (Abschnitt 2) und anschließend anhand einer beispielhaften Analyse zum Einfluss der Komplexität (Variation der Anzahl an Dimensionen) illustriert (Abschnitt 3). Hierfür bildet eine experimentelle Online-Vignettenstudie zur Einkommensgerechtigkeit die Datenbasis. Der Beitrag schließt mit einem Ausblick auf anknüpfende Forschungsfragen (Abschnitt 4).

2 Forschungsprogramm

Das übergreifende Ziel des Forschungsprojektes besteht in der Erarbeitung von anwendungsbezogenen Kriterien für die Durchführung von faktoriellen Surveys, wobei der Schwerpunkt auf der Einstellungsmessung in allgemeinen Bevölkerungsumfragen liegt. In mehreren Experimentalreihen werden methodische Aspekte variiert, um den Einfluss von Designmerkmalen auf das Antwortverhalten und die Datenqualität zu studieren. Grob lassen sich dabei die zwei Aspekte „Komplexität" und „Darstellungsweise" unterscheiden.

Zur *Komplexität* zählt zunächst die Frage, wie viele Dimensionen idealerweise verwendet werden. In der Einleitung wurde bereits eine Vignette präsentiert, wie sie im vorliegenden Forschungsprojekt eingesetzt wird. Es handelt sich um eine Beschreibung mittels fünf Dimensionen. Diese wenigen Informationen sind möglicherweise für ein fundiertes Gerechtigkeitsurteil unzureichend. Die Befragten müssten sich dann fehlende Informationen selbst konstruieren (Wason et al. 2002: 52; DeShazo/Fermo 2002: 131f.). So könnten sie in dem Beispiel aufgrund des Alters von 50 Jahren und einem Hochschulabschluss eine ca. 25-jährige Berufserfahrung assoziieren. Möglicherweise wird aufgrund der Tatsache, dass es sich um eine Frau handelt, zugleich eine familienbedingte Erwerbsunterbrechung unterstellt und damit eine weitaus geringere Berufserfahrung. Unterschiede in der Bewertung weiblicher und männlicher Einkommen entsprächen dann nicht mehr zwangsläufig einer Präferenz für geschlechtsspezifische Entlohnungen, sondern können ebenso Folge unterschiedlicher Annahmen zu unbekannten Merkmale sein (theoretisch wäre von einer statistischen Diskriminierung zu sprechen). Wie dieses Beispiel zeigt, geht bei sehr inhaltsleeren Vignetten der experimentelle Vorteil einer hohen Kontrolle über das Untersuchungsdesign verloren. Es sind Drittvariableneffekte nicht mehr auszuschließen. Aus diesem Grund und auch im Hinblick auf einen höheren Informationsgewinn sind höhere Anzahlen an Dimensionen empfehlenswert.

Ab wann sind es aber zu viele? Mehr Merkmale bedeuten längere Fallbeispiele und mehr Informationen, die simultan in ein Urteil integriert werden müssen. Ab irgendeinem Punkt ist eine kognitive Überforderung der Befragten zu erwarten. Diskutierte Symptome sind Befragungsabbrüche und inkonsistentere Urteile. Zudem dürften speziell sehr komplexe Vorgaben zu Rückgriffen auf vereinfachende Heuristiken motivieren, bei denen die Urteile nur mehr auf wenige zentrale Merkmale gestützt werden (Auspurg et al. 2009a). Welche dieser Folgen eintritt ist ebenso unklar wie die Dimensionszahl, ab der mit ihnen zu rechnen ist. Bislang vorherrschend sind sechs bis acht Dimensionen (vgl. Beck/Opp 2001: 287), diese Richtgröße ist allerdings ein reiner Erfahrungswert und keinesfalls empirisch validiert. Im vorliegenden Projekt werden daher unterschiedliche Dimensionszahlen gegeneinander getestet (fünf, acht und zwölf Dimensionen).

Ähnlich stellt sich die Frage nach einer idealen Anzahl an Vignetten, die einzelnen Befragten vorgelegt werden. Eine höhere Anzahl an Vignetten pro Befragten wird aufgrund der damit erreichten Vergrößerung der Fallzahl (welche sich als Produkt aus der Anzahl an Befragten und der Anzahl an Urteilen pro Befragten ergibt) empfohlen und aufgrund der Möglichkeit, Urteilsregeln befragtenspezifisch bestimmen zu können (Jasso 2006).[1] Hinzu kommt, dass die Einflussgewichte einzelner Merkmale grundsätzlich umso präziser schätzbar sind, je mehr Vignetten aus dem so genannten „Vignettenuniversum" zum Einsatz kommen. Als dieses werden alle möglichen Merkmalskombinationen bzw. Vignetten bezeichnet, sein Umfang errechnet sich als kartesisches Produkt der Ausprägungen der einzelnen Dimensionen und erreicht rasch Fallzahlen, die für eine Bewertung durch einzelne Be-

1 Die Zahl der Urteile muss hierfür die Zahl der Merkmale bzw. Regressionskoeffizienten überschreiten.

fragte zu hoch sind.[2] Der Ausweg besteht in der Verwendung von Stichproben, wobei zur besseren Ausschöpfung des Universums üblicherweise mehrere unterschiedliche Fragebogenversionen bzw. „Decks“ (auch als „Sets“ bezeichnet) eingesetzt werden. Nach der vorliegenden Methodenforschung sind bewusste fraktionalisierte Auswahlen, die sich an statistischen Kriterien wie einem Erhalt einer möglichst hohen Unkorreliertheit und Varianz der Dimensionen orientieren, den alternativen Zufallsstichproben aufgrund ihres höheren Informationsgehalts vorzuziehen (Steiner/ Atzmüller 2006; Dülmer 2007). Unabhängig von dem Stichprobenverfahren gilt es die Anzahl zu ziehender Vignetten und damit die Anzahl unterschiedlicher Decks sowie Vignetten pro Befragten festzulegen. Wünschenswerte Eigenschaften des Universums (wie vor allem die Unabhängigkeit der Dimensionen) werden grundsätzlich umso besser bewahrt, je höher die Fallzahl an Vignetten ist. Dies ist gegen mögliche Ermüdungs- und Langeweile-Effekte bei den Befragten abzuwägen. Die mehrfache Bewertungsaufgabe durch einzelne Befragte gilt zudem wegen möglichen Reihenfolge- und Ausstrahlungseffekten (*carry-over*-Effekten) als problematisch. Hinzu tritt zumindest bei geschlossenen Antwortskalen das Risiko von Antwortzensierungen. Haben die Befragten bereits eine Extremkategorie ausgewählt und werden sie dann mit einer noch ungerechteren Vignette konfrontiert, können sie ihr Urteil nicht mehr adäquat abstufen (es sei denn, sie haben die Möglichkeit, das frühere Urteil zu korrigieren und scheuen diese Mühe nicht). Schließlich steht die sequentielle Bearbeitung im Verdacht, ein sozial erwünschtes Antwortverhalten zu provozieren. Wenn die Befragten merken, welche Merkmale variieren, können sie ihre Antworten im Hinblick auf diese „politisch korrekt“ abstimmen – etwa ihre Gerechtigkeitsurteile in Bezug auf Frauen und Männer bewusst angleichen, um dem Verdacht einer Einkommensdiskriminierung von Frauen zu begegnen (Jann 2003). Diesbezüglich erscheint es am sinnvollsten, nur eine Vignette pro Befragten zu verwenden. Wiederum gilt es also, zur Bestimmung der idealen Vorgehensweise diverse methodische Aspekte gegeneinander abzuwägen. Das vorliegende Projekt setzt hierzu Splits mit 10, 20 und 30 Vignetten ein.[3]

Als weiterer Aspekt von Komplexität wird schließlich geprüft, ob der für ähnliche Erhebungsverfahren (Conjoint-Analysen und Choice-Experimente) gut belegte *number-of-levels*-Effekt bei faktoriellen Suveys ebenso auftritt (z.B. Wittink et al. 1982, 1990; Perrey 1996): Ziehen Merkmale inhaltsunabhängig eine stärkere Aufmerksamkeit auf sich, weil sie mit vergleichsweise mehr Levels variieren? Um dies zu testen, werden in einem Split die Ausprägungen des Alters verdoppelt (es werden Zwischenkategorien eingefügt).

Im Hinblick auf den *Präsentationsmodus* der Vignetten interessieren zunächst die Effekte unterschiedlicher Antwortskalen. Zur Verhinderung der angesprochenen Antwortzensierungen werden in der Praxis offene Skalen, insbesondere Magnitude-Skalen empfohlen (z.B. Liebig/Mau 2002, 2005). Eine offene Frage ist allerdings, ob solche weniger gängigen Antwortformate nicht kognitive Überforderungen provozieren, speziell bei älteren und weniger gebildeten Befragten (Schaeffer/Bradburn 1989 für allgemeine Evidenzen). Teil des Forschungsprogramms ist es daher, Magnitude- gegen Ra-

2 Bei dem eingangs präsentierten Beispiel mit fünf Dimensionen besteht das Design aus zwei Ausprägungen des Geschlechts (Mann/Frau), acht Alterskategorien (25/30/35/…/60 Jahre), drei Kategorien für den Abschluss (kein Abschluss/abgeschlossene Berufsausbildung/Hochschulabschluss), sowie jeweils zehn Berufs- und Einkommenskategorien. Damit ergibt sich eine Universum von 2x8x3x10x10 = 4.800 möglichen Vignetten

3 Die in der Literatur zu findenden Richtlinien (z.B. zehn bis zwanzig Vignetten pro Befragten in dem Einführungsaufsatz von Opp und Beck 2001) sind wiederum keinesfalls empirisch gestützt und es finden sich durchaus auch Beispiele mit weitaus höheren Anzahlen, ohne dass nennenswerte Probleme berichtet würden. Die Maximalzahl liegt bei ganzen 95 Vignetten pro Befragten (Studie von Rossi et al. 1974).

ting-Skalen zu testen. Daneben wird geprüft, ob sich Antwortzensierungen reduzieren lassen, wenn die Befragten zu Beginn mit Extremfällen konfrontiert werden. Diese bewusste Reihung der Vignetten wird einer rein zufälligen Sortierung gegenübergestellt, welche wiederum zur Vermeidung von Ausstrahlungs- und Reihenfolgeeffekten empfohlen wird (Rossi/ Anderson 1982). Weiter wird untersucht, welche Effekte sich durch eine tabellarische statt Fließtext-Formulierung der Vignetten ergeben. Eine tabellarische Auflistung der Dimensionen verspricht eine bessere Übersichtlichkeit für die Befragten und könnte damit gerade bei sehr komplexen Vignetten zu sicheren und konsistenteren Urteilen verhelfen. Für Fließtextvignetten wird dagegen argumentiert, dass sie das Hineinversetzen in die Situationen erleichtern. Die Befragten würden „intuitiver" antworten und aufgrund der indirekteren Urteilsaufgabe weniger Anreizen zu einem sozial erwünschten Antwortverhalten erliegen (Alexander/Becker 1978). Wiederum fehlen empirische Hinweise dafür, welche Variante empfehlenswerter ist. Schließlich interessiert, inwieweit *Range*-Effekte auftreten, also die Befragten ihr Antwortverhalten an die vorgegebenen Spannweiten der Dimensionen adaptieren (für Conjoint-Analysen und Choice-Experimente: Creyer/Ross 1988; Ohler et al. 2000). Führen Einkommensbeträge zu anderen Urteilen, wenn sie als Maximalwert in den Vignetten eines Befragten herausstechen, oder werden die Urteile unabhängig vom präsentierten Wertebereich gefällt? Inwieweit gilt also der für andere Frageformate bekannte Effekt, dass vorgegebene Kategorien den Befragten als Referenzpunkte für ihre Antworten dienen (dazu z.B. Diekmann 2007)?

Die beschriebenen Methodenexperimente werden innerhalb der beiden Gruppen (Komplexität und Präsentationsmodus) allesamt nochmals vollständig untereinander gekreuzt, um mögliche Wechselwirkungen beobachten zu können. Beispielsweise dürften Lern- und Ermüdungseffekte insbesondere dann auftreten, wenn hohe Anzahlen an Dimensionen verwendet und den einzelnen Befragten zugleich viele Urteile abverlangt werden. Ähnlich sind Antwortzensierungen insbesondere dann zu erwarten, wenn die Befragten viele Urteile nacheinander in eine Antwortskala einzupassen haben. Insgesamt ergeben sich durch diese Kreuzung mehrere Duzend experimentelle Splits, die mit zwei Haupterhebungen umgesetzt werden (Auspurg/Wehrli 2009; Auspurg et al 2009b für Details). *Erstens* wird die komplette Experimentalreihe an einer möglichst homogenen Befragtenstichprobe getestet, um die Wirkung der Designmerkmale möglichst unabhängig von Merkmalen der Befragten beobachten zu können. Die dazu durchgeführte Online-Befragung von gut 2.000 Studierenden der Sozialwissenschaften ist bereits erfolgreich abgeschlossen (ein Ausschnitt aus dieser wird in Abschnitt 3 für die Prüfung der Hypothesen zur Komplexität genutzt). In dieser Studierendenbefragung wurde ein Teil der Probanden auch in einer Panelstudie zu drei Zeitpunkten mit exakt denselben Vignetten konfrontiert, um Aufschlüsse über die zeitliche Stabilität bzw. Reliabilität ihrer Urteile zu gewinnen.[4] *Zweitens* wird ein Ausschnitt der Experimente zusätzlich an einer Stichprobe von N = 1.500 der Allgemeinbevölkerung umgesetzt. Es handelt sich um die Experimente zur Komplexität, bei denen vornehmlich Wechselwirkungen mit Merkmalen der Befragten (wie ihrer kognitiven Belastbarkeit) anzunehmen sind. Dieser fast abgeschlossene Erhebungsteil bezweckt speziell eine Prüfung der Durchführbarkeit in allgemeinen Bevölkerungsumfragen, wozu zusätzlich eine Variation des Befragungsmodus vorgenommen wird (Einsatz von Computer Assisted Personal Interviews [CAPI] vs. telefonische Rekrutierung mit anschließendem Versand eines elektronischen Fragebogens [Computer Assisted Self Interviewing, CASI] oder Papier-Fragebogens [Paper And Pencil Interview, PAPI)].

4 Hierbei wurde zwischen den Befragten die Komplexität (Anzahl an Vignetten und Dimensionen) variiert.

Diese zusätzlichen Splits sollen zur Klärung beitragen, ob das spezielle Antwortformat von Vignetten die Anwesenheit eines assistierenden Interviewers erfordert.[5]

Zielgrößen bilden stets die Gerechtigkeitsurteile und deren Abstufung bzw. Varianz sowie die Antwortkonsistenz. Weiter interessiert, wie viele Dimensionen die Befragten in ihre Urteile einbinden und inwieweit sie vereinfachende Heuristiken anwenden (Mazzotta/Opaluch 1995; Gigerenzer/Goldstein 1996; Lloyd 2003). Automatisch erfasst werden schließlich Metadaten wie die Antwortgeschwindigkeit, ein mögliches Zurückblättern im Fragebogen (als Indikator für Antwortkorrekturen) und (Item-)Nonresponses. Mit diesem umfangreichen Datenmaterial lassen sich die angesprochenen methodischen Fragen analysieren und zusätzlich modelltheoretische Fragen der Datenauswertung faktorieller Surveys sowie einige ungeklärte Stichprobenaspekte bearbeiten. Im Folgenden wird die Vorgehensweise anhand erster Analysen zu den Auswirkungen unterschiedlich komplexer Vignetten (Variation der Anzahl an Dimensionen) auf die Gerechtigkeitsurteile illustriert.

3 Empirische Analysen zur Komplexität

3.1 Forschungsstand und Hypothesen

Mögliche Folgen einer zu geringen oder hohen Anzahl an variablen Dimensionen wurden bereits angesprochen. Grob lassen sich ein Rückgriff auf Heuristiken und Inkonsistenzen im Antwortverhalten unterscheiden. Erstere stehen in enger Verbindung zu der von Kritikern zuweilen vorgebrachten Vermutung, Befragte würden sich im Sinne eines sozial erwünschten Antwortverhaltens primär auf konsistente Urteile konzentrieren (z.B. Faia 1980). Die Methodenforschung zu den verwandten Conjoint-Analysen und Choice-Experimenten zeigt, dass Befragte eine hohe Konsistenz trotz hoher Komplexität mit geringer Anstrengung durch die Fokussierung auf wenige zentrale Merkmale erreichen. Erst wenn sich zentrale Urteilskriterien nicht zwischen den Fallbeispielen unterscheiden, werden von den Befragten zusätzliche Merkmale beachtet (zu solchen lexikographischen Entscheidungsregeln z.B. Payne et al. 1993).[6] Trifft ein solcher Rückgriff auf vereinfachende Heuristiken ähnlich bei Vignettenstudien zu, sollte sich dies darin manifestieren, dass einzelne Dimensionen in komplexen Settings (in denen ein Umschwenken auf Heuristiken lohnt) eine geringere Urteilsrelevanz erhalten als in weniger komplexen Settings (in denen sich noch alle Merkmale mit geringer Anstrengung beachten lassen). Ab welcher Dimensionszahl ist aber von einer hohen Komplexität zu sprechen? Bislang dominieren in Vignettenstudien etwa sieben Dimensionen (Beck/Opp 2001). Diese Anzahl gilt auch aufgrund von kognitionswissenschaftlichen Befunden als sinnvolle Obergrenze, nach denen das menschliche Kurzzeitgedächtnis etwa sieben Informationseinheiten problemlos speichern kann (Zimbardo 1988: 275). Allerdings ist fraglich, ob diese Regel wirklich auf Vignetten

5 Zusätzlich zu diesen beiden Erhebungsteilen konnte die Projektgruppe Vignetten in der SOEP-Pretesterhebung 2008 einsetzen. Bei diesen Vignetten werden zwar keine experimentellen Variationen vorgenommen, aufgrund der umfangreichen Dokumentation des empfundenen Schwierigkeitsgrades seitens der Befragten und Interviewer gewähren diese Daten gleichwohl weiteren Aufschluss über methodische Aspekte (dazu Sauer et al. 2009).

6 Dies entspräche auch dem aus einer Rational-Choice-Perspektive ableitbaren Prinzip einer möglichst effizienten Kosten-Nutzenbilanz: „the choice of a decision strategy is governed by a desire to make a good decision while at the same time minimizing cognitive effort" (Lloyd 2003: 398).

anwendbar ist – es geht hier ja weniger um das Memorieren einzelner Informationen, als vielmehr die Beurteilung einer zusammenhängenden Personen- oder Situationsbeschreibung. Es soll daher zunächst folgende Annahme geprüft werden:

H_1: Im Fall von mehr als sieben Dimensionen werden einzelne Merkmale stärker ausgeblendet und sind daher weniger urteilsrelevant als im Falle von bis zu sieben Dimensionen.

Die zweite diskutierte Folge einer hohen Komplexität ist eine kognitive Überforderung, welche sich in einem inkonsistenten Antwortverhalten äußert (Rossi/Anderson 1982:59). Für eine geringe Konsistenz wird aber ebenso ein Informationsmangel als Ursache angenommen. Wenn sich die Befragten fehlende Informationen individuell „zusammen reimen" müssen, bedeutet dies technisch eine geringe Standardisierung bzw. hohe unbeobachtete Heterogenität.[7] Treffen beide Prozesse zu, ist ein umgekehrt u-förmiger Zusammenhang zwischen der Anzahl an Dimensionen und Antwortkonsistenz zu erwarten, wobei als Richtwert für den Wendepunkt die bereits genannte Zahl von etwa sieben Dimensionen herangezogen werden kann. Als Hypothese formuliert:

H_2: Das Antwortverhalten ist bei etwa sieben Dimensionen konsistenter als bei deutlich weniger oder mehr Dimensionen.

3.2 Datengrundlage

Als Datengrundlage dient ein Ausschnitt aus der Onlinebefragung von Studierenden der Sozialwissenschaften. Deren Teilnehmer wurden im Sommersemester 2008 über Methodenveranstaltungen im Grundstudium der Sozialwissenschaften an insgesamt 25 Universitäten in ganz Deutschland rekrutiert (für Einzelheiten: Auspurg et al. 2009b).[8] Aus den mit diesen CASI-Befragungen (Computer Assisted Self Interviewing) gewonnenen Daten wird die Experimentalreihe zur Anzahl an Dimensionen herausgegriffen. Die Dimensionszahl wird dort zwischen den Befragten mit fünf, acht oder zwölf Dimensionen variiert. Die Maximalzahl von zwölf überschreitet die bislang in der Literatur vorherrschende Anzahl von etwa sieben Dimensionen deutlich. Speziell bei dieser Variante sollten sich also die postulierten Anzeichen hoher Komplexität (Ausblenden von Merkmalen und/oder geringere Antwortkonsistenz) zeigen. Die maximale Antwortkonsistenz ist für den Split mit acht Di-

7 Vorgeschlagen wird ebenso, dass zu wenige Merkmalsvorgaben kognitiv belastend sind, weil es mit weniger Variationen schwieriger ist, Unterschiede in den Fallbeispielen zu erkennen und damit zwischen ihnen zu differenzieren (für dieses Argument bei Choice-Experimenten Hensher 2006a). Als ein erster Beleg für einen solchen „Information-Underload" lassen sich Befunde einer Wiederholungsbefragung werten, bei der Studierende zu drei Messzeitpunkten mit den jeweils selben Vignetten befragt wurden: Die Stabilität der Urteile erwies sich bei acht Dimensionen höher als bei fünf Dimensionen (Liebig/Meyermann/Schulze 2006). Bis auf diese erste Pilot-Studie fehlen allerdings empirische Hinweise.

8 Etwa zur Hälfte führten die Studierenden die Befragung direkt während den Veranstaltungszeiten in einem CIP-Pool und unter Aufsicht eines Projektmitarbeiters durch. Von der anderen Hälfte der Befragungsteilnehmer wurde in den Veranstaltungen lediglich eine E-Mail-Adresse erhoben, um ihnen anschließend den Link zur Befragung zuzustellen. Das erste Verfahren ist wegen der höheren Standardisierung vorzuziehen, war aber aus technischen oder praktischen Gründen nicht in allen Veranstaltungen möglich. In beiden Varianten wurde jeweils das Methodenexperiment nicht erwähnt und stattdessen ausschließlich ein inhaltliches Interesse an Einkommensgerechtigkeit genannt, um eventuelle Verzerrungen des Antwortverhaltens durch eine Orientierung der Befragten an den von ihnen erahnten Forschungshypothesen zu verhindern (sog. Reaktivität bzw. Hawthorne-Effekt, dazu allgemein Diekmann 2007).

mensionen anzunehmen, während für den Split mit fünf Dimensionen aufgrund der Informationsarmut zumindest befragtenübergreifend eine geringere Urteilskonsistenz zu erwarten ist. Eine Beispielvignette für den Split mit fünf Dimensionen wurde bereits eingangs präsentiert (Abbildung 1 in Abschnitt 1). Die achtdimensionalen Vignetten beinhalten zusätzlich die drei Merkmale Berufserfahrung, Dauer der Betriebszugehörigkeit und Anzahl Kinder; in der zwölfdimensionalen Variante werden noch ergänzend Informationen zum Gesundheitszustand und der Leistung der Vignettenperson sowie der ökonomischen Situation und Größe des Betriebs vorgegeben. Aufgabe der Befragten ist es stets, die Gerechtigkeit des Bruttoeinkommens auf einer elfstufigen Ratingskala einzustufen. Sie werden zufällig auf die einzelnen Splits zugewiesen und beantworten zehn bis 30 Vignetten.[9] In allen Splits wird dabei dieselbe fraktionalisierte Auswahl von 240 Vignetten eingesetzt. Dies gewährleistet, dass sich die Splits in den Varianzen und Korrelationen der Vignettenmerkmale entsprechen und somit Unterschiede in Schätzparametern (wie Regressionskoeffizienten) tatsächlich Folge eines anderen Urteilverhaltens sind (Unterschiede in der inhaltlichen Komposition der Vignetten oder in der statistischen Güte ihrer Stichproben sind als Ursachen ausgeschlossen).[10]

3.3 Ergebnisse

Bevor die Hypothesen mit Hilfe von multivariaten Verfahren getestet werden, interessieren ein paar deskriptive Angaben zum Rücklauf. Die beschriebenen Splits haben insgesamt 952 Befragte bearbeitet und dabei etwa 15.500 Vignettenurteile abgegeben. Pro Experimentalzelle liegen mindestens 5.000 Urteile vor (vgl. Tabelle 1). Die Fallzahlen sind sehr ausgewogen, womit sich statistische Parameter wie die R^2-Werte in Regressionen gut zwischen den Splits vergleichen lassen. Finden sich Unterschiede, sind diese zumindest nicht fallzahlenbedingt.

Tabelle 1: Anzahlen an Urteilen und Befragten nach experimentellem Split

	5 Dimensionen	8 Dimensionen	12 Dimensionen	Gesamt
Anzahl Urteile	5.093	5.114	5.535	15.512
Anzahl Befragte	305	321	326	952

9 Die Anzahl an Vignetten wird unabhängig von der Anzahl an Dimensionen variiert. Im Mittel liegen von den einzelnen Befragten 16 Urteile vor.

10 Praktisch wurde dies erreicht, indem eine fraktionalisierte Auswahl von 240 Vignetten für die Variante mit zwölf Dimensionen gebildet wurde (orientiert an dem Kriterium einer maximalen D-Effizienz, dazu allgemein Kuhfeld 2005; speziell für die Umsetzung in diesem Projekt Auspurg/Wehrli 2009; Sauer et al. 2009). Für die Splits mit nur fünf oder acht Dimensionen wurden dann die dort irrelevanten Dimensionen gelöscht. Zwar ließen sich für diese Splits weitaus effizientere Stichproben bilden (also solche mit einer höheren Unkorreliertheit und Varianz der Merkmale), gerade solche *statistischen* Designmerkmale sollten aber konstant gehalten werden um die reinen *methodischen* Effekte beobachten zu können (dazu ausführlicher Auspurg et al. 2009a). Für die folgenden Auswertungen werden zudem ausschließlich Splits mit einer befraggtenspezifischen Zufallsreihenfolge von Vignetten herangezogen. Die einzelnen Vignetten verteilen sich aufgrund der Randomisierung gleichmäßig auf die Bearbeitungspositionen, womit Reihenfolge bzw. *carry-over*-Effekte neutralisiert werden (deren Inzidenz andernfalls mit dem Split variieren könnte).

Abbildung 1 präsentiert die Verteilung der Urteile, aufgeschlüsselt für die Splits mit fünf, acht oder zwölf Dimensionen. Deutlich wird eine Häufung auf dem Wert Null („gerecht“). Zudem ist speziell die untere Extremkategorie („ungerechterweise zu niedrig“) vergleichsweise stark besetzt. Sollte es sich hierbei um Antwortzensierungen handeln (Urteile, bei denen die Befragten gerne einen noch extremeren Wert angekreuzt hätten), werden ohne spezielle Korrekturverfahren (wie z.B. Tobit-Regressionen) die Einflüsse der unabhängigen Variablen tendenziell unterschätzt (Wooldridge 2003; Berk 1983).[11] Da die Anzeichen für Zensierungen jedoch insgesamt gering sind und zur Kontrolle durchgeführte Tobit-Regressionen zu den gleichen Schlussfolgerungen führen, werden im Folgenden einfache

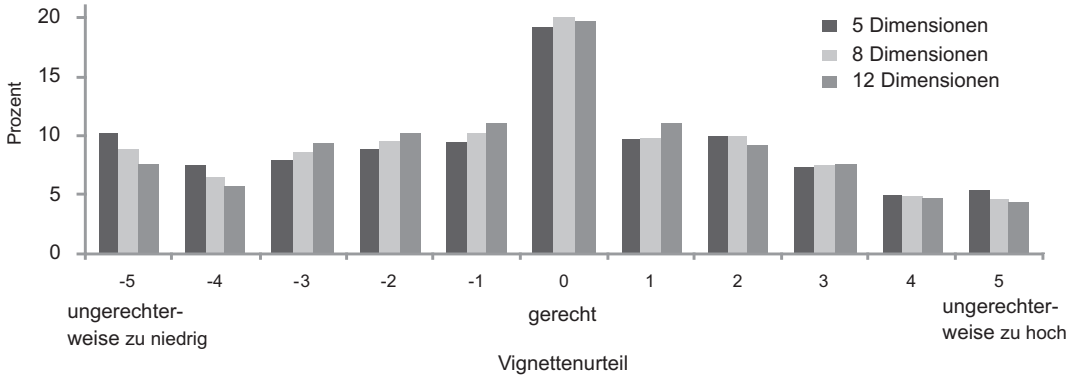

Abbildung 2: Verteilung der Urteile für unterschiedlich komplexe Vignetten

Welche Auswirkungen hat nun die Anzahl der Dimensionen auf das Antwortverhalten? Aufschlüsse hierzu liefern die in Tabelle 2 aufgeführten Regressionen. Aufgrund der hierarchischen Datenstruktur (dazu Hox et al. 1991; Auspurg et al. 2009a) werden diese jeweils mit robusten Standardfehlern geschätzt.[12] In den Modellen 1, 2a und 3a sind jeweils nur die grundständigen, in allen Splits auftretenden Dimensionen als erklärende Variablen einbezogen. Diese Modelle entsprechen sich somit in den (Anzahlen der) zu schätzenden Parametern, was bessere Vergleichsmöglichkeiten eröffnet. Um Drittvariableneffekte ausschließen zu können (welche aufgrund der fraktionalisierten Vignettenauswahl allerdings kaum auftreten dürften), werden für die Splits mit acht und zwölf Dimensionen zusätzlich Modelle mit allen Dimensionen berechnet (Modelle 2b und 3b).[13]

11 Diese Verzerrung wäre dann in dem Split mit fünf Dimensionen etwas ausgeprägter (die Antwortballung tritt bei diesem am stärksten auf).

12 Von einzelnen Befragten liegen mehrere Urteile vor, womit die Unabhängigkeitsannahme verletzt ist bzw. es zu Autokorrelationen der Fehlerterme kommt.

13 Aufgrund des vorgenommenen Ausschlusses sehr unplausibler Kombinationen (etwa Kombinationen von Berufen, die einen Hochschulabschluss voraussetzen, mit der Ausbildungskategorie „kein Abschluss“) oder völlig unlogischer Kombinationen (wie „keine Berufserfahrung“ und „lange Betriebszugehörigkeit“) treten trotz der fraktionalisierten Auswahl geringe Korrelationen der Merkmale auf.

Tabelle 2: OLS-Regressionen der Vignettenurteile für unterschiedlich komplexe Vignetten (Koeffizienten; in Klammern robuste Standardfehler)

	Modell 1 5 Dimensionen	Modell 2a 8 Dimensionen	Modell 2b 8 Dimensionen m. Kontrollvar.	Modell 3a 12 Dimensionen	Modell 3b 12 Dimensionen m. Kontrollvar.
Geschlecht (1=Frau)	-0.101*	-0.076	-0.087*	0.029	0.020
	(0.057)	(0.053)	(0.052)	(0.053)	(0.051)
Alter [10 Jahre]	-0.089***	-0.053*	-0.028	- 0.074***	-0.063**
	(0.027)	(0.028)	(0.027)	(0.027)	(0.025)
Bildung (Ref.: Ohne Abschluss)					
▪ Ausbildungsabschl.	-0.739***	-0.486***	-0.488***	-0.386***	-0.382***
	(0.072)	(0.062)	(0.057)	(0.062)	(0.056)
▪ Hochschulabschl.	-1.158***	-0.749***	-0.734***	-0.664***	-0.639***
	(0.082)	(0.070)	(0.067)	(0.064)	(0.061)
Prestige [10 MPS]	-0.152***	-0.126***	-0.127***	-0.107***	-0.108***
	(0.009)	(0.007)	(0.008)	(0.007)	(0.007)
Einkommen [1000 €]	0.505***	0.481***	0.484***	0.459***	0.462***
	(0.011)	(0.010)	(0.010)	(0.009)	(0.009)
Kontrolle für					
▪ Anzahl Kinder			X		X
▪ Berufserfahrung			X		X
▪ Dauer Betriebszugeh.			X		X
▪ Betriebsgröße					X
▪ Wirtschl. Lage Betrieb					X
▪ Gesundheitszustand					X
▪ Leistung					X
Konstante	-0.149	-0.648***	0.070	-0.812***	0.272
	(0.165)	(0.166)	(0.159)	(0.151)	(0.163)
Beobachtungen	5093	5114	5114	5305	5305
(Befragte)	(306)	(321)	(321)	(327)	(327)
R^2	0.494	0.480	0.505	0.471	0.503
Adj. R^2	0.494	0.480	0.504	0.470	0,501

*** p<0.01, ** p<0.05, * p<0.1; in Klammern robuste Standardfehler (Huber-White-Korrektur)

Kurz zur inhaltlichen Lesart der Ergebnisse. Bei der vorliegenden Kodierung der abhängigen Variablen bedeuten positive (negative) Koeffizientenwerte, dass das Einkommen als ungerechterweise zu hoch (niedrig) empfunden wird. Negative Effekte lassen sich somit als eine Erhöhung des als gerecht empfundenen Bruttoeinkommens bei einem Anstieg metrischer Variablen deuten (oder bei Wechsel einer Dummy-Variablen von der Referenz- auf die angegebene Kategorie). Nach allen drei Modellen wird somit älteren Personen ein höheres Einkommen zugestanden als jüngeren und für Personen mit

Ausbildungsabschluss ein höheres Einkommen als gerecht empfunden als für vergleichbare Personen ohne Abschluss. Für unser methodisches Forschungsinteresse interessanter ist allerdings, ob sich die Koeffizientenwerte zwischen den Modellen unterscheiden.

Dies ist in der Tat der Fall. Vergleicht man zunächst die mit denselben Variablen geschätzten Modelle 1, 2a und 3a untereinander, stellt man fest, dass die Koeffizienten für den Split mit fünf Dimensionen stets einen etwas höheren Betrag aufweisen als für den Split mit acht und speziell zwölf Dimensionen. Besonders deutlich wird dies beim Hochschulabschluss, dieser Koeffizient weist bei fünf Dimensionen einen fast doppelt so hohen Betrag auf als in dem zwölfdimensionalen Split (Koeffizientenwert von -1,158 bei fünf Dimensionen; -0,734 bei acht; -0,638 bei zwölf Dimensionen). Die Unterschiede zwischen fünf (Modell 1) und zwölf Dimensionen (Modell 3a) erweisen sich dabei größtenteils als statistisch signifikant.[14] Wie ein Vergleich der Modelle 2a mit 2b sowie 3a mit 3b zeigt, ist die Abschwächung der Effekte mit steigender Dimensionszahl auch kein Drittvariableneffekt der bei den höher-dimensionalen Varianten hinzukommenden Merkmale. Die Effekte bleiben durchweg stabil, wenn für diese zusätzlichen Merkmale kontrolliert wird (lediglich der Einfluss des Alters schwächt sich im Split für acht Dimensionen dann geringfügig ab).

Die Effekte der weiteren Dimensionen, welche hier aus Platzgründen nicht ausgewiesen sind, erweisen sich ebenfalls mehrheitlich als signifikant und durchweg als plausibel. Zusammengenommen sprechen diese Beobachtungen dafür, dass „the best piece of information“ von der Zusammenstellung der Dimensionen abhängt. Treten Dimensionen mit mehreren anderen zusammen, sind sie tendenziell weniger urteilsrelevant als wenn sie eine von wenigen Merkmalsdimensionen darstellen. Dies deckt sich mit der ersten Hypothese (abnehmende Bedeutung von Merkmalen, wenn sieben Dimensionen überschritten werden). Für eine Feststellung, ob sich dahinter kognitive Überforderungen und/oder Rückgriffe auf Heuristiken verbergen, wären allerdings noch vertiefende Analysen anzustellen.[15]

Unabhängig davon, welche Ursache nun konkret zutrifft, ist festzuhalten, dass die Einflussstärke von Vignettenmerkmalen mitunter eine Funktion ihrer Einzelständigkeit ist. Die Signifikanzen erweisen sich als relativ robust gegenüber diesem Effekt,[16] was im Hinblick auf die Testung von Hypothesen zum grundsätzlichen Einfluss von Merkmalen schon einmal eine gute Nachricht ist. Zurückhaltend interpretiert werden sollten jedoch die absoluten Effektstärken, sie sind nach den vorliegenden Befunden nur für ähnlich komplexe Vignetten (und möglicherweise ähnlich kognitiv belastbare Befragte) miteinander vergleichbar. Sofern sich die Effekte der verschiedenen Dimensionen aber stets um den gleichen Faktor abschwächen, wären zumindest noch Relationen zwischen einzelnen Koeffi-

14 Prüfung mittels eines Chow-Test: F [6, 953] = 5,31; p= 0,000. Einzeln getestet erweisen sich die Effekte der Bildungsabschlüsse, des Einkommens und des Prestige als signifikant verschieden. Erläuterungen zum Chow-Tests finden sich z.B. in Wooldridge 2003; für eine Anwendung bei Vignetten: Auspurg et al. 2009a.

15 Die Anzeichen für Antwortzensierungen sind wie dargelegt gering und zudem bei dem Split mit fünf Dimensionen am stärksten ausgeprägt. Dennoch erscheint es sinnvoll, die hier vorgenommenen Analysen nochmals mit einer offenen Antwortskala zu validieren (was mit den im Projekt gesammelten Daten prinzipiell möglich ist). Zudem wäre zu prüfen, ob die für andere Verfahren vorgeschlagenen Analysestrategien zur Feststellung von Heuristiken auf Vignettenanalysen übertragbar sind.

16 Allerdings verfehlt das Geschlecht im Split mit zwölf Dimensionen anders als in den beiden anderen das Signifikanzniveau.

zienten verlässlich interpretierbar.[17] Dies ist etwa deshalb relevant, da Relationen zwischen Koeffizienten den sehr informativen Berechnungen gerechter Einkommensunterschiede zu Grunde liegen. Deren Vergleichbarkeit über die Splits soll hier noch kurz am Beispiel des just gender wage gaps geprüft werden. Nach leichten Umformungen der OLS-Schätzgleichungen errechnet sich dieser folgendermaßen[18]:

$$\text{Just gender wage gap} = \frac{\beta_{Geschlecht}}{\beta_{Einkommen}} * 1000 \text{ Euro}$$

Während die Befragten in den Splits mit fünf und acht Dimensionen weiblichen Erwerbstätigen im Mittel (knapp) 200 Euro weniger Bruttomonatsgehalt zugestehen als vergleichbaren männlichen Erwerbstätigen (vgl. Tabelle 3), dreht sich der *just gender wage gap* in dem Split mit zwölf Dimensionen um zu einer geringfügigen, nicht signifikanten Überbezahlung von weiblichen Einkommensbeziehern um gut 40 Euro (der Regressionskoeffizient für das Geschlechts ist nicht signifikant von Null verschieden, vgl. Tabelle 2). Ob diese Variationen nach Split eher einer Überforderung mit der hohen Informationsfülle in den zwölfdimensionalen Vignetten geschuldet ist oder aber einer geschlechtsspezifischen Unterstellung einkommensrelevanter Merkmale aufgrund der Informationsarmut in den weniger-dimensionalen Vignetten, bleibt weiterführenden Analysen vorbehalten. In diesen wäre zudem zu prüfen, ob sich ähnliche Effekte bei offenen Antwortskalen zeigen (wie sie bislang primär zur Bestimmung von just wage gaps eingesetzt werden).[19] Das Beispiel veranschaulicht gleichwohl, dass Designeffekte unbedingt bei den Ergebnisinterpretationen zu beachten sind.

Tabelle 3: Just gender wage gap nach Anzahl Dimensionen

Anzahl Dimensionen	$\beta_{Geschlecht}$	$\beta_{Einkommen}$	just gender wage gap (gerechtes Gehalt Frauen minus Männer)
5 Dimensionen	- 0.101	0.505	- 200 Euro
8 Dimensionen	- 0.087	0.481	- 179 Euro
12 Dimensionen	0.020	0.462	+ 43 Euro

Anmerkung: Schätzungen mit Einbezug aller Dimensionen (Modelle 1,2b, und 3b in Tabelle 2)

Abschließend ist noch die Hypothese zur Antwortkonsistenz zu prüfen (postuliert wurde, dass diese bei etwa sieben Dimensionen am höchsten ist, vgl. H_2). Als Indikator für diese ist es in Vignettenstu-

17 Dies deshalb, weil sich eine Abschwächung der Koeffizienten um den gleichen Faktor bei einer Verhältnissetzung heraus kürzt. Dies wird auch an nachfolgendem Beispiel ersichtlich. Eine solche gleichmäßige Abschwächung um den gleichen Faktor wäre bei Überforderungen, kaum aber bei Heuristiken zu erwarten.

18 Der Hintergrund ist der, dass männliche und weibliche Einkommensbezieher in den Vignetten das gleiche Urteil erhalten, wenn sich die Terme $\beta_{Geschlecht}$ und - $\beta_{Einkommen}$*1000 Euro die Waage halten. Errechnet wird, mit wie viel Euro Einkommensabschlag die Tatsache, dass es sich um eine weibliche statt männliche Person handelt, im Hinblick auf das Gerechtigkeitsurteil kompensiert wird.

19 Zudem erweisen sich die Befunde möglicherweise gegenüber diesem Designeffekt als robuster, wenn andere Modellierungen verwendet werden. Zu prüfen wäre etwa, ob Regressionsschätzungen mit logarithmiertem Einkommen eine bessere Spezifikation bieten. Oftmals werden aufgrund einer Ableitung aus Gerechtigkeitstheorien leicht abgewandelte Formeln verwendet und zudem bei Magnitude-Skalen befragtenspezifische Regressionsschätzungen als Berechnungsgrundlage verwendet (z.B. Liebig/Mau 2002, 2005). An dem Grundprinzip einer Verhältnissetzung der Koeffizienten ändert dies jedoch nichts.

dien üblich, die aufgeklärte Varianz bzw. R^2-Werte heranzuziehen (Rossi/Anderson 1982: 21 ff.; Jasso 2006). Für den hier interessierenden Modellvergleich bieten aufgrund der unterschiedlichen Anzahlen an unabhängigen Variablen, damit Freiheitsgraden, allerdings die adjustierten R^2-Werte eine bessere Basis.[20] Diese sind ebenfalls in Tabelle 2 aufgeführt (letzte Zeile). Entgegen der Annahme unterscheidet sich die Modellgüte bzw. Antwortkonsistenz kaum zwischen den einzelnen Splits, sie liegt in allen Modellen nahe bei dem Wert von 0,5. An diesem Indikator beurteilt, erweisen sich also selbst die Vignetten mit zwölf Dimensionen nicht als zu komplex für die Befragten. Geringe R^2-Werte in gepoolten Regressionen können allerdings neben einer geringen Antwortkonsistenz ebenfalls nicht modellierte Unterschiede in den Gerechtigkeitsprinzipien der Befragten widerspiegeln. Zur Absicherung sollen daher noch die R^2-Werte aus *befragtenspezifischen* Regressionen betrachtet werden (Schätzungen getrennt über die Urteile der einzelnen Befragten). Deren Mittelwerte sind in Tabelle 4 aufgeführt. Einbezogen wurden dort allein Befragte mit mindestens 30 gültigen Urteilen, da geringere Fallzahlen keine stabilen Schätzungen zulassen.[21] Die Antwortkonsistenz (erneut gemessen an den adjustierten R^2-Werten) erweist sich wiederum in allen Splits als sehr ähnlich. Lediglich der Maximalwert sticht etwas hervor, welcher überraschenderweise in der Modellschätzung mit allen zwölf Dimensionen erreicht wird (Modell 3b). Nach obigen Analysen ist dieser Befund damit zu erklären, dass sich die Befragten bei hoher Komplexität lediglich auf einen Ausschnitt der für sie besonders zentralen Informationen konzentrieren – mit anderen Worten sich die höhere Komplexität eigenständig reduzieren. Diese Interpretationen wäre noch mit weiteren Analysen zu untermauern (in denen etwa gezielt heuristische Entscheidungsmuster modelliert werden, dazu Payne et al. 1993). Zudem erscheinen Replikationen mit offenen Antwortskalen sinnvoll, die den Befragten mehr Raum für die Abstufung ihrer Urteile einräumen.

Tabelle 4: Modellgüte befragtenspezifischer Regressionen

	Modell 1 5 Dimensionen	Modell 2a 8 Dimensionen	Modell 2b 8 Dimensionen mit Kontrollvar.	Modell 3a 12 Dimensionen	Modell 3b 12 Dimensionen mit Kontrollvar.
N Befragte	60	49	49	56	56
Mittelwert adj. R^2	0,57	0,56	0,58	0,55	0,63
Standardabw.	(0,17)	(0,15)	(0,16)	(0,15)	(0,18)

20 Die R^2-Werte steigen unweigerlich mit Anzahl an Dimensionen an. Bei der Berechnung adjustierter R^2-Werte wird dies berücksichtigt, indem die Anzahl an Variablen bzw. Freiheitsgraden mit in die Berechnung einfließt (z.B. Wooldridge 2003). Allerdings sind diese Korrekturformeln nicht unumstritten (der Hauptkritikpunkt ist, dass die Interpretierbarkeit als Anteil erklärter Varianz verloren geht). Gleichwohl bilden sie gängige Maßzahlen für die Modellgüte und werden daher auch hier herangezogen.

21 Die Anzahl an Urteilen pro Befragten ist bei befragtenspezifischen Regressionen schließlich gleichbedeutend mit der Fallzahl. Zumindest bei Einbezug von zwölf Dimensionen lassen sich aufgrund der hohen Variablenzahl hinreichend stabile Schätzungen erst ab 30 Urteilen pro Befragten erzielen (zu den statistischen Grundlagen Hox et al. 1991).

4 Fazit und Ausblick

Die hier präsentierten Analysen zum Einfluss der Anzahl an Dimensionen verdeutlichen, dass methodische Effekte bei der Interpretation und Konstruktion faktorieller Surveys zu beachten sind. Die Einflussstärken von Vignettenmerkmalen sind mitunter eine Funktion ihrer Einzelständigkeit und ihrer inhaltlichen Komposition. Dieser Effekt ist speziell dann zu berücksichtigen, wenn unterschiedlich komplexe Designs miteinander verglichen werden sollen oder detaillierte Aussagen interessieren – etwa in Form der hier demonstrierten Berechnungen gerechter Lohnunterschiede. Bevor die Ursachen dieser Unterschiede nicht genauer geklärt und in Auswertungen kontrollierbar sind, kann der Rat nur lauten, derartige Berechnungen nicht „auf den Cent genau" auszulegen. Die positive Nachricht ist allerdings, dass sich die grundsätzlichen Einflüsse von Merkmalen (Signifikanzwerte und Vorzeichen) als relativ robust erweisen. Für viele Anwendungsziele (etwa Tests von Theorien) dürfte dieser Informationsgehalt bereits ausreichend sein.

Zugleich wurde weiterer Forschungsbedarf offensichtlich. In zusätzlichen Analysen wäre zu klären, wodurch die Fokussierung der Befragten auf wenige Dimensionen geleitet wird. Ist es wirklich „the best piece of information" auf das sie sich bei komplexeren Vignetten konzentrieren? Oder sind es möglicherweise (auch) die am stärksten auffallenden, da am stärksten variierenden Merkmale, auf welche sie ihre Urteile vordergründig stützen? Für die verwandten Conjoint-Analysen ist schließlich bereits gut belegt, dass Merkmale *inhaltsunabhängig* umso mehr beachtetet werden, mit umso mehr Ausprägungen und einer umso größeren Spannweite sie variieren (z.B. Creyer/Ross 1988; Wittink et al. 1990; Ohler et al. 2000). Für Vignettenstudien stehen Nachweise solcher *number-of-levels*- und *range*-Effekten noch aus. Sollte die Fokussierung auf wenige Merkmale bei hoher Komplexität Symptom einer kognitiven Überforderung sein,[22] sind derartige Effekte speziell bei kognitiv weniger belastbaren Befragten zu erwarten. Bildungs- und altersspezifische Urteilsmuster wären dann nicht mehr reiner Ausdruck subgruppenspezifischer Einstellungen, sondern (auch) methodisch bedingt. Um Fehlschlüssen vorzubeugen, bedarf es gerade hier weiterer Methodenforschung.[23]

Mit den im beschriebenen Forschungsprojekt erhobenen Daten lassen sich derartige Aspekte untersuchen. Sie beinhalten zusätzliche Indikatoren für die Antwortstrategien (etwa Antwortzeiten und andere Metadaten) und erlauben es überdies mit der Bevölkerungsumfrage, das Urteilsverhalten unterschiedlicher Bevölkerungsgruppen zu studieren. Speziell mit den Paneldaten wird sich zudem Aufschluss über die Reliabilität der Urteile als zentraler Vorbedingung für deren Validität gewinnen lassen. All dies trägt zur Prüfung und Weiterentwicklung eines Verfahrens bei, dass sich international für verschiedene sozialwissenschaftliche Forschungsziele etabliert hat. Viele Befunde zu einem optimalen Ausmaß an Komplexität, idealen Antwort- und Präsentationsformaten erscheinen aber ebenso über das spezielle Format von Vignettenstudien hinaus von allgemeiner Bedeutung.

22 Die tendenzielle Ausblendung von einzelnen Merkmalen bei hoher Komplexität kann alternativ auch Folge eines generellen Bestrebens nach Reduktion von Komplexität sein, wie es schließlich auch für viele alltägliche Urteile und Entscheidungen gültig ist. Reduktionen von Komplexität oder Heuristiken bedeuten daher nicht zwangsläufig invalide Messungen, möglicherweise korrespondieren sie sogar besser mit realen Entscheidungsmustern (Swait/Adamowitz 2001:147; Hensher 2006b).

23 Schließlich dürften kognitive Überlastungen auch von der Anzahl an abverlangten Urteilen abhängen, es empfiehlt sich also zusätzlich Wechselwirkungen mit der Anzahl an Vignetten zu testen.

Literatur

Alexander, Cheryl S./Becker, Henry Jay (1978): The Use of Vignettes in Survey Research. Public Opinion Quarterly 42: 93-104.

Auspurg, Katrin/Hinz, Thomas/Liebig, Stefan (2009a): Komplexität von Vignetten, Lerneffekte und Plausibilität im Faktoriellen Survey. Methoden – Daten – Analysen 3: 59-96.

Auspurg, Katrin, Thomas Hinz, Stefan Liebig und Carsten Sauer (2009b): Feldbericht der Studie „Einkommensgerechtigkeit in Deutschland". Technical Report # 2 des DFG-Projekts Forschungsprojekts „Der faktorielle Survey als Instrument zur Einstellungsmessung in Umfragen. Bielfeld/Konstanz. Universität Bielefeld/Universität Konstanz.

Auspurg, Katrin/Wehrli, Stefan (2009): Codebuch und Dokumentation. Einkommensgerechtigkeit in Deutschland. Universität Konstanz/ ETH Zürich.

Beck, Michael/Opp, Karl Dieter (2001): Der faktorielle Survey und die Messung von Normen. Kölner Zeitschrift für Soziologie und Sozialpsychologie 53: 283-306.

Berk, Richard A. (1983): An Introduction to Sample Selection Bias in Sociological Data. American Sociological Review 48: 386-398.

Creyer, Elizabeth/Ross, William T. (1988): The Effect of Range-Frequency Manipulations on Conjoint Importance Weight Stability. Advances in Consumer Research 15: 505-509.

DeShazo, J.R./Fermo, German (2002): Designing Choice Sets for Stated Preference Methods: The Effects of Complexity on Choice Consistency. Journal of Environmental Economics and Management 44: 123-143.

Diekmann, Andreas (2007): Empirische Sozialforschung. Grundlagen, Methoden, Anwendungen. Reinbek bei Hamburg: Rohwolt.

Dülmer, Hermann (2007): Experimental Plans in Factorial Surveys: Random or Quota Design? Sociological Methods & Research 35: 382-409.

Faia, Michael A. (1980): The Vagaries of the Vignette World: A Comment on Alves and Rossi. American Journal of Sociology 85: 951-954.

Gigerenzer, Gerd/Goldstein, Daniel G. (1996): Reasoning the Fast and Frugal Way: Models of Bounded Rationality. Psychological Review 103: 650-669.

Hensher, David A. (2006a): Revealing Differences in Willingness to Pay due to the Dimensionality of Stated Choice Designs: An Initial Assessment. Environmental & Resource Economics 34: 7-44.

Hensher, David. A. (2006b): How do Respondents Process Stated Choice Experiments? Atribute Consideration under Varying Information Load. Journal of Applied Econometrics 21: 861-878.

Hox, Joop J./ Kreft, Ita G./Hermkens, Piet L.J. (1991): The Analysis of Factorial Surveys. Sociological Methods & Research 19: 493-510.

Jann, Ben (2003): Lohngerechtigkeit und Geschlechterdiskriminierung: Experimentelle Evidenz. Unveröffentlichtes Manuskript an der Eidgenössischen Technischen Hochschule Zürich.

Jasso, Guillermina (2006): Factorial Survey Methods for Studying Beliefs and Judgments. Sociological Methods Research 34: 334-423.

Kuhfeld, Warren F. (2005): Marketing Research Methods in SAS. Experimental Design, Choice, Conjoint and Graphical Techniques. Cary: SAS Institute.

Liebig, Stefan/Mau, Steffen (2002): Einstellungen zur sozialen Mindestsicherung. Ein Vorschlag zur differenzierten Erfassung normativer Urteile. Kölner Zeitschrift für Soziologie und Sozialpsychologie 54: 109-134.

Liebig, Stefan/Mau, Steffen (2005): Wann ist ein Steuersystem gerecht? Zeitschrift für Soziologie 34: 468-491.

Liebig, Stefan/ Meyermann, Alexia/Schulze, Andrea (2006): Temporal stability of justice evaluations. Paper presented at the 11th Conference of the International Society for Justice Research. Berlin: Humboldt Universität.

Lloyd, Andrew J. (2003): Threats to the estimation of benefit: are preference Elicitation methods accurate? Health Economics 12: 393-402.

Lorenz, Wilhelm (1993): Diskriminierung. S. 119-144 in: Ramb, Bernd-Thomas und Manfred Tietzel (Hg.): Ökonomische Verhaltenstheorie. München: Vahlen.

Mazzotta, Marisa J./Opaluch, James J. (1995): Decision Making when Choices are Complex. A Test of Heiner's Hyotheses. Land Economics 71: 500-515.

Ohler, Tobias/Le, Aihong/Louviere, Jordan/Swait, Joffre D. (2000): Attribute Range Effects in Binary Response Tasks. Marketing Letters 11: 249-260.

Payne, J., J. Bettman, und E. Johnson, 1993: The Adaptive Decision Maker. Cambridge University Press: Cambridge, MA.

Perrey, Jesko (1996): Erhebungsdesign-Effekte bei der Conjoint-Analyse. Marketing – Zeitschrift für Forschung und Praxis 18: 105-116.

Rossi, Peter H./Anderson, Andy B. (1982): The Factorial Survey Approach: An Introduction. S. 15-67 in: Rossi/Nock (Hg.): Measuring Social Judgements. The Factorial Survey Approach. Beverly Hills u.a.: Sage.

Rossi, Peter H./Sampson, William A./Ch. E. Bose, Christine E./ Jasso, Guillermina/Passel Jeff (1974): Measuring Household Social Standing. Social Science Research 3: 169-190.

Sauer, Carsten/Auspurg, Katrin/Hinz, Thomas/Liebig, Stefan/Schupp, Jürgen (2009): Die Bewertung von Erwerbseinkommen – Methodische und inhaltliche Analysen zu einer Vignettenstudie im Rahmen des SOEP-Pretest 2008. DIW Data Documentation 42. Berlin: Deutsches Institut für Wirtschaftsforschung (DIW).

Schaeffer, Nora Cate/Bradburn, Norman M. (1989): Respondent Behavior in Magnitude Estimation. In: Journal of the American Statistical Association 84: 402-413.

Steiner, Peter M./Atzmüller, Christiane (2006): Experimentelle Vignettendesigns in Faktoriellen Surveys. Kölner Zeitschrift für Soziologie und Sozialpsychologie 58: 117-146.

Swait, Joffre/Adamowicz, Wiktor (2001): The Influence of Task Complexity on Consumer Choice: A Latent Class Model of Decision Strategy Switching. Journal of Consumer Research 28: 135-148.

Wallander, Lisa (2009): 25 years of factorial surveys in sociology: a review. Social Science Research 38: 505-520.

Wason, Kelly D./Polonsky, Michael. J./Hyman, Michael R. (2002): Designing Vignette Studies in Marketing. In: Australasian Marketing Journal 10: 41-58.

Wittink, Dick. R./Krishnamurthi, Laksham/Nutter, Julia B. (1982): Comparing derived importance weights across attributes. Journal of Consumer Research 8: 471-474.

Wittink, Dick. R./Krishnamurthi, Laksham/Reibstein, David J. (1990): The Effect of Differences in the Number of Attribute Levels on Conjoint Results. Marketing Letters 1: 113-123.

Wooldridge, Jeffrey M. (2003): Introductory Econometrics. A Modern Approach. Mason, Ohio: Thomson.

Zimbardo, Philip G. (1988): Psychologie. Berlin u.a.: Springer.

Zu den Autoren

Auspurg, Katrin, Diplom-Soziologin, wissenschaftliche Mitarbeiterin am Fachbereich Soziologie der Universität Konstanz. Nach einem Abschluss als Diplom-Sozialpädagogin Studium der Soziologie und Statistik an der Ludwig-Maximilians-Universität München, Forschungsschwerpunkte: Soziale Ungleichheit, Diskriminierung, quantitative Sozialforschung
E-Mail: Katrin.Auspurg@uni-konstanz.de

Hinz, Thomas, Prof. Dr., Professor für Empirische Sozialforschung an der Universität Konstanz, Studium der Soziologie, Sozialpsychologie und Statistik sowie Promotion und Habilitation an der LMU München, Gastprofessor an der ETH Zürich und Yale University, Forschungsschwerpunkte: Methoden der empirischen Sozialforschung, Arbeitsmarkt- und Wirtschaftssoziologie
E-Mail: Thomas.Hinz@uni-konstanz.de

Liebig, Stefan, Prof. Dr., Professor für Soziale Ungleichheit und Sozialstrukturanalyse an der Fakultät für Soziologie der Universität Bielefeld, Forschungsprofessor am DIW Berlin, Studium der Soziologie und Ev. Theologie, Promotion an der HU Berlin, Habilitation an der LMU München. Forschungsschwerpunkte: Soziale Ungleichheit und Gerechtigkeit, Methoden der empirischen Sozialforschung
E-Mail: stefan.liebig@uni-bielefeld.de

Sauer, Carsten, M.A., wissenschaftlicher Mitarbeiter an der Fakultät für Soziologie der Universität Bielefeld, Studium der Soziologie, Philosophie und Volkswirtschaftslehre an der Universität Konstanz. Forschungsschwerpunkte: Soziale Ungleichheit und Gerechtigkeit, Methoden der empirischen Sozialforschung
E-Mail: carsten.sauer@uni-bielefeld.de